This Page Is Inserted by IFW Operations
and is not a part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

W 1409

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 09-006799
(43)Date of publication of application : 10.01.1997

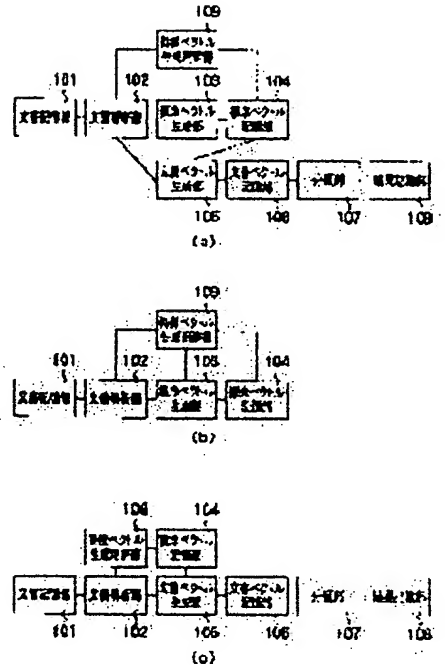
(51)Int.Cl. G06F 17/30
G06F 17/27

(21)Application number : 07-151640
(22)Date of filing : 19.06.1995

(71)Applicant : SHARP CORP
(72)Inventor : YUASA NATSUKI

(54) DOCUMENT SORTING DEVICE AND DOCUMENT RETRIEVING DEVICE

(57)Abstract:
PURPOSE: To provide a device with which a document can be sorted or retrieved regardlessly of the kind of language.
CONSTITUTION: The document sorting device is provided with a document storage part 101 for storing document data, document analytic part 102 for analyzing the document data, conceptional vector generating part 103 for generating the feature vectors of conceptional identifiers in the document, conceptional vector storage part 104 for storing those feature vectors, document vector generating part 105 for generating the feature vectors of the document from the feature vectors of conceptional identifiers contained in the document, document vector storage part 106 for storing those feature vectors, sorting part 107 for sorting the document while utilizing the degree of similarity between the document feature vectors, result storage part 108 for storing the sorted result, and dictionary 109 for feature vector generation registering words or conceptional identifiers to be used when generating the feature vectors.



LEGAL STATUS

[Date of request for examination] 18.12.1998
[Date of sending the examiner's decision of rejection] 29.02.2000
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]
[Patent number]
[Date of registration]
[Number of appeal against examiner's decision of rejection]
[Date of requesting appeal against examiner's decision of rejection]
[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報**（Ａ）　(11)特許出願公開番号

特開平9－6799

(43)公開日　平成９年(1997)１月10日

(51)Int.Cl.[6]		識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ	17/30		9289－5L	Ｇ０６Ｆ	15/403	３５０Ｚ
	17/27		8420－5L		15/38	Ｍ
			9289－5L		15/40	３７０Ａ

審査請求　未請求　請求項の数６　ＯＬ　（全 19 頁）

(21)出願番号　特願平7－151640

(22)出願日　平成７年(1995)６月19日

(71)出願人　000005049
シャープ株式会社
大阪府大阪市阿倍野区長池町22番22号

(72)発明者　湯浅　夏樹
大阪府大阪市阿倍野区長池町22番22号　シャープ株式会社内

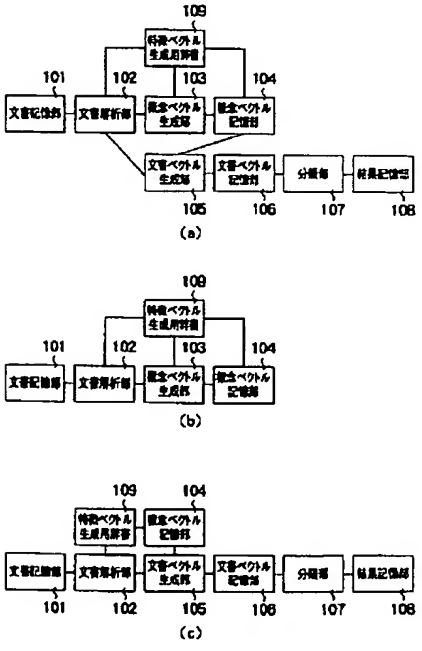
(74)代理人　弁理士　川口　義雄　（外１名）

(54)【発明の名称】　文書分類装置及び文書検索装置

(57)【要約】

【目的】　言語の種類を問わず、文書を分類したり検索したりすることができる装置を提供する。

【構成】　文書分類装置において、文書データを記憶する文書記憶部１０１と、文書データを解析する文書解析部１０２と、文書中の概念識別子の特徴ベクトルを生成する概念ベクトル生成部１０３と、その特徴ベクトルを記憶する概念ベクトル記憶部１０４と、文書内に含まれている概念識別子の特徴ベクトルから文書の特徴ベクトルを生成する文書ベクトル生成部１０５と、その特徴ベクトルを記憶する文書ベクトル記憶部１０６と、文書特徴ベクトル間の類似度を利用して文書を分類する分類部１０７と、その分類した結果を記憶する結果記憶部１０８と、特徴ベクトル生成時に使用する単語や概念識別子が登録されている特徴ベクトル生成用辞書１０９とを備える。



【特許請求の範囲】
【請求項１】　文書の内容にしたがって文書の分類を行う文書分類装置であって、
文書データを記憶する文書記憶部と、
予め定められた単語及び概念識別子を登録した特徴ベクトル生成用辞書と、
前記特徴ベクトル生成用辞書によって、記憶した文書データの単語を解析する文書解析部と、
前記特徴ベクトル生成用辞書によって、文書データの単語を概念識別子に変換し、概念識別子間の共起関係に基づいて、概念識別子の特徴ベクトルを自動的に生成する概念ベクトル生成部と、
生成した概念識別子の特徴ベクトルを記憶する概念ベクトル記憶部と、
概念識別子の特徴ベクトルから文書の特徴ベクトルを生成する文書ベクトル生成部と、
文書の特徴ベクトルを記憶する文書ベクトル記憶部と、
文書の特徴ベクトル間の類似度を利用して文書を分類する分類部と、
分類した結果を記憶する結果記憶部と、を含むことを特徴とする文書分類装置。
【請求項２】　前記結果記憶部に記憶された分類ごとに概念識別子の出現率を調べ、分類に有用な概念識別子を選出し、分類に有用な概念識別子を前記特徴ベクトル生成用辞書に登録する、有用概念識別子選出部をさらに含み、分類に有用な概念識別子を用いることで分類の精度を向上させることを特徴とする請求項１に記載の文書分類装置。
【請求項３】　前記結果記憶部に記憶された分類ごとに、概念識別子の特徴ベクトルと文書の特徴ベクトルとの少なくとも一方を用いて、その分類を代表する文書の特徴ベクトルを求める代表ベクトル生成部と、分類を代表する文書の特徴ベクトルを記憶する代表ベクトル記憶部とをさらに含む請求項１又は請求項２に記載の文書分類装置。
【請求項４】　前記特徴ベクトル生成用辞書が複数の言語の辞書を含んでおり、前記複数の言語のどの言語の単語であっても同じ概念の単語は同じ概念識別子に変換し、言語の種類によらない文書分類を行う請求項１から請求項３のいずれか一項に記載の文書分類装置。
【請求項５】　文書検索装置であって、
文書データを記憶する文書記憶部と、
検索文を入力する検索文入力部と、
予め定められた単語及び概念識別子を登録した特徴ベクトル生成用辞書と、
前記特徴ベクトル生成用辞書によって、記憶した文書データの単語を解析する文書解析部と、
前記特徴ベクトル生成用辞書によって、文書データの単語を概念識別子に変換し、概念識別子間の共起関係に基づいて、概念識別子の特徴ベクトルを自動的に生成する概念ベクトル生成部と、
概念識別子の特徴ベクトルを記憶する概念ベクトル記憶部と、
文書データ及び検索文中に含まれている概念識別子の特徴ベクトルから文書データ及び検索文の特徴ベクトルを生成する文書ベクトル生成部と、
文書データ及び前記検索文の特徴ベクトルを記憶する文書ベクトル記憶部と、
文書データの特徴ベクトルと検索文の特徴ベクトルとの類似度を利用して文書データ中から検索文に類似した文を検索する検索部と、
その検索した結果を出力する出力部と、を含む文書検索装置。
【請求項６】　前記特徴ベクトル生成用辞書が複数の言語の辞書を含んでおり、前記複数の言語のどの言語の単語であっても同じ概念の単語は同じ概念識別子に変換し、言語の種類によらない文書検索を行う請求項５に記載の文書検索装置。
【発明の詳細な説明】
【０００１】
【産業上の利用分野】本発明は、文書や電子メール等を分類する文書分類装置及び大量の文書データの中から必要な情報を取り出す文書検索装置に関する。本発明の装置は、ワープロやファイリングシステムなどの分野にも利用される。さらに、どの言語も区別することなく使用することができる。
【０００２】
【従来の技術】文書を自動的に分類する方法としては、例えば、田村他「統計的手法による文書自動分類」（情報処理学会第３６回全国大会論文集，１９８７年）や、特開平２－１５８８７１号公報に開示された「文書分類装置」や、特開平６－３４８７５５号公報に開示された「文書分類方法およびそのシステム」があげられる。田村他の方法はキーワードの出現頻度の分野による偏りを示す指標としてカイ自乗値を求め文書を分類するものであるが、この方法は、キーワードの出現頻度の偏りを用いるために、予め標本データを人手によって分野別に分類してカイ自乗値を計算し、分類用データを用意しておく必要がある。他方、特開平２－１５８８７１号公報に開示された「文書分類装置」は、標本データを分野別に分類しておく必要はないが、文書分類用のシソーラスやキーワード分類項目を予め人手により登録しておく必要がある。また、特開平６－３４８７５５号公報に開示された「文書分類方法およびそのシステム」では文書分類用のシソーラスやキーワード分類項目を登録しておく必要はないが、一分類が一文書データ以上からなる分類済みの文書データを用意しておく必要がある。
【０００３】
【発明が解決しようとする課題】従来の文書分類方法では、標本データを人手によって分野別に分類したり、シ

ソーラスやキーワード分類項目を人手により登録しておく必要があり、分類に際してなんらかの人手を必要とするという問題があった。特開平６－３４８７５５号公報に開示された「文書分類方法およびそのシステム」においては、一分類が一文書データでよいので、人手による手間は比較的少ないが、分類精度を向上させたければより多くの分類済み文書データを用意しておく必要がある。

【０００４】また、従来の文書分類方式では同一言語での（日本語なら日本語の）文書を分類することしか考慮されていなかった。

【０００５】本発明の課題は、上記問題を解決するために、特に分類されていない状態の単語データや文書データを用意するだけで、文書データ内での出現頻度から分類に用いる特徴ベクトルを自動的に生成し、この特徴ベクトルを用いることで、未知の文書を自動的に分類する装置を提供することである。単語データや文書データは特に分類されていない状態で使用するので、人手による手間を全く必要としない。さらに、本発明の副次的な課題は、言語によらない分類を行なうことができる装置を提供することである。

【０００６】

【課題を解決するための手段】請求項１に記載の発明の文書分類装置は、文書データを記憶する文書記憶部と、予め定められた単語及び概念識別子を登録した特徴ベクトル生成用辞書と、特徴ベクトル生成用辞書によって、記憶した文書データの単語を解析する文書解析部と、特徴ベクトル生成用辞書によって、文書データの単語を概念識別子に変換し、概念識別子間の共起関係に基づいて、概念識別子の特徴ベクトルを自動的に生成する概念ベクトル生成部と、生成した概念識別子の特徴ベクトルを記憶する概念ベクトル記憶部と、概念識別子の特徴ベクトルから文書の特徴ベクトルを生成する文書ベクトル生成部と、文書の特徴ベクトルを記憶する文書ベクトル記憶部と、文書の特徴ベクトル間の類似度を利用して文書を分類する分類部と、分類した結果を記憶する結果記憶部とを含むことを特徴とする。

【０００７】請求項２に記載の発明の文書分類装置は、結果記憶部に記憶された分類ごとに概念識別子の出現率を調べ、分類に有用な概念識別子を選出し、分類に有用な概念識別子を特徴ベクトル生成用辞書に登録する、有用概念識別部をさらに含み、分類に有用な概念識別子を用いることで分類の精度を向上させることを特徴とする。

【０００８】請求項３に記載の発明の文書分類装置は、結果記憶部に記憶された分類ごとに、概念識別子の特徴ベクトルと文書の特徴ベクトルとの少なくとも一方を用いて、その分類を代表する文書の特徴ベクトルを求める代表ベクトル生成部と、分類を代表する文書の特徴ベクトルを記憶する代表ベクトル記憶部とをさらに含むことを特徴とする。

【０００９】請求項４に記載の発明の文書分類装置は、特徴ベクトル生成用辞書が複数の言語の辞書を含んでおり、複数の言語のどの言語の単語であっても同じ概念の単語は同じ概念識別子に変換し、言語の種類によらない文書分類を行うことを特徴とする。

【００１０】請求項５に記載の文書検索装置は、文書データを記憶する文書記憶部と、検索文を入力する検索文入力部と、予め定められた単語及び概念識別子を登録した特徴ベクトル生成用辞書と、特徴ベクトル生成用辞書によって、記憶した文書データの単語を解析する文書解析部と、特徴ベクトル生成用辞書によって、文書データの単語を概念識別子に変換し、概念識別子間の共起関係に基づいて、概念識別子の特徴ベクトルを自動的に生成する概念ベクトル生成部と、概念識別子の特徴ベクトルを記憶する概念ベクトル記憶部と、文書データ及び検索文中に含まれている概念識別子の特徴ベクトルから文書データ及び検索文の特徴ベクトルを生成する文書ベクトル生成部と、文書データ及び前記検索文の特徴ベクトルを記憶する文書ベクトル記憶部と、文書データの特徴ベクトルと検索文の特徴ベクトルとの類似度を利用して文書データ中から検索文に類似した文を検索する検索部と、その検索した結果を出力する出力部とを含むことを特徴とする。

【００１１】請求項６に記載の文書検索装置は、特徴ベクトル生成用辞書が複数の言語の辞書を含んでおり、複数の言語のどの言語の単語であっても同じ概念の単語は同じ概念識別子に変換し、言語の種類によらない文書検索を行うことを特徴とする。

【００１２】

【作用】請求項１に記載の文書分類装置においては、文書の学習と学習に基づいた文書の分類が行われる。文書の学習においては、文書記憶部に記憶されている学習対象の文書データの内容が文書解析部に渡され、特徴ベクトル生成用辞書の単語を使用して文書の解析が行われる。つぎに、概念ベクトル生成部において、特徴ベクトル生成用辞書の概念識別子を使用して単語から概念識別子への変換が行われ、概念識別子間の共起関係を用いて概念識別子の特徴を表現する概念識別子の特徴ベクトルが自動的に生成される。こうして生成された概念識別子の特徴ベクトルは、概念ベクトル記憶部に記憶される。文書の分類においては、文書記憶部に記憶されている分類対象の文書データの内容が文書解析部に渡され、特徴ベクトル生成用辞書の単語を使用して文書の解析が行われる。つぎに、文書ベクトル生成部において、概念ベクトル記憶部に登録された概念識別子から、文書の特徴ベクトルを生成する。こうして生成された文書の特徴ベクトルは、文書ベクトル記憶部に記憶される。分類部において、文書の特徴ベクトルの類似度によって文書が分類される。分類結果は、結果記憶部に記憶される。

【００１３】請求項２に記載の文書分類装置においては、結果記憶部に記憶された分類ごとに概念識別子の出現率を調べ、分類に有用な概念識別子を選出し、分類に有用な概念識別子を特徴ベクトル生成用辞書に登録する、有用概念識別部をさらに含むように構成されているので、分類に有用な概念識別子を用いることによって、特徴ベクトルの記憶空間を削減したり、分類の精度を向上させることができる。

【００１４】請求項３に記載の文書分類装置においては、結果記憶部に記憶された分類ごとに、概念識別子や文書の特徴ベクトルを用いて、その分類を代表する文書の特徴ベクトルを求める代表ベクトル生成部と、分類を代表する文書の特徴ベクトルを記憶する代表ベクトル記憶部とをさらに含むように構成されているので、一度各分類群の代表ベクトルを生成してしまえば、新たな文書データを分類するときには、その文書の特徴ベクトルと各分類群の代表ベクトルとの比較を行なうだけでその文書がどの分類群に属すかを判定できるようになる。

【００１５】請求項４に記載の文書分類装置においては、特徴ベクトル生成用辞書が複数の言語の辞書を含んでおり、複数の言語のどの言語の単語であっても同じ概念の単語は同じ概念識別子に変換し、言語の種類によらない文書分類を行うことができる。

【００１６】請求項５に記載の文書検索装置においては、文書の学習と学習に基づいた文書の検索が行われる。文書の学習においては、文書記憶部に記憶されている学習対象の文書データの内容が文書解析部に渡され、特徴ベクトル生成用辞書の単語を使用して文書の解析が行われる。つぎに、概念ベクトル生成部において、特徴ベクトル生成用辞書の概念識別子を使用して単語から概念識別子への変換が行われ、概念識別子間の共起関係を用いて概念識別子の特徴を表現する概念識別子の特徴ベクトルが自動的に生成される。こうして生成された概念識別子の特徴ベクトルは、概念ベクトル記憶部に記憶される。文書の検索においては、検索文入力部から検索キーとなる文書が入力され、文書解析部に渡され、特徴ベクトル生成用辞書の単語を使用して文書の解析が行われる。つぎに、文書ベクトル生成部において、概念ベクトル記憶部に登録された概念識別子から、文書の特徴ベクトルを生成する。こうして生成された文書の特徴ベクトルは、文書ベクトル記憶部に記憶される。検索部において、検索キーとなる文書と学習された文書との特徴ベクトルの類似度が比較され、類似度の高いものが検索結果として出力部に渡され、検索結果として出力される。

【００１７】請求項６に記載の文書検索装置においては、特徴ベクトル生成用辞書が複数の言語の辞書を含んでおり、複数の言語のどの言語の単語であっても同じ概念の単語は同じ概念識別子に変換し、言語の種類によらない文書検索を行うことができる。

【００１８】

【実施例】請求項１に記載の発明の文書分類装置の一実施例を図１に示す。ここで、図１（ａ）は、全体の装置構成、図１（ｂ）は、学習時に使用される装置の構成、図１（ｃ）は、分類時に使用される装置の構成を夫々示す。図中１０１は文書記憶部、１０２は文書解析部、１０３は概念ベクトル生成部、１０４は概念ベクトル記憶部、１０５は文書ベクトル生成部、１０６は文書ベクトル記憶部、１０７は分類部、１０８は結果記憶部、１０９は特徴ベクトル生成用辞書である。

【００１９】文書記憶部１０１には、学習に用いるための文書や、分類する文書を記憶する。文書解析部１０２は文書記憶部１０１から文書を渡され、特徴ベクトル生成用辞書１０９中の単語辞書を用いてその文書の形態素解析を行なう。ここで、文書の形態素解析とは、文書を単語等に分けることをいう。

【００２０】概念ベクトルを学習する場合の各構成要素の作用の概要について、図１（ｂ）に基づいて説明する。概念ベクトル生成部１０３では、文書解析部１０２から渡された単語データを、特徴ベクトル生成用辞書１０９中の概念辞書（単語と概念識別子との関連付けを行なっている辞書）を参照して概念識別子に変換し、概念識別子間の共起関係を利用して概念識別子の特徴ベクトルを生成する。概念ベクトル記憶部１０４は、概念ベクトル生成部１０３で生成された概念識別子の特徴ベクトルを記憶する。

【００２１】つぎに、文書を分類する場合の各構成要素の作用の概要について、図１（ｃ）に基づいて説明する。文書ベクトル生成部１０５では、文書解析部１０２から渡された単語データを、特徴ベクトル生成用辞書１０９中の概念辞書を参照して概念識別子に変換し、そこで得られた概念識別子の特徴ベクトルを概念ベクトル記憶部１０４を参照して求め、文書中から得られる全ての単語についてこのようにして求めた概念識別子の特徴ベクトルから（平均化するなどして）文書の特徴ベクトルを求める。文書ベクトル記憶部１０６は、文書ベクトル生成部で求められた文書の特徴ベクトルを記憶する。分類部１０７は、文書ベクトル記憶部１０６から渡された文書の特徴ベクトルを用いて、文書を分類する。結果記憶部１０８は、分類部１０７で分類された文書の情報（どの文書がどの分野に分類されたか）を記憶する。

【００２２】特徴ベクトル生成用辞書１０９は、文書を形態素解析する時に用いる単語辞書と、各単語に関連付けられた概念識別子を求めるための概念辞書とからなる。これは必ずしも二つの辞書に分けられているということではなく、一つの辞書において、各単語に概念識別子が割り当てられているような辞書であってもよい。

【００２３】一般に通常の文書に使用されている全ての単語に関連付けられた概念識別子の個数を合計すると非常に大きな数値になるため、特徴ベクトルを作成する際に用いる概念識別子の個数を制限しておくのが好まし

い。このために特徴ベクトル生成用辞書１０９の概念辞書において、ここに登録されている概念識別子のみを用いて概念識別子の特徴ベクトルを作成することで、特徴ベクトルの記憶空間の巨大化を抑えることができる。

【００２４】概念識別子の特徴ベクトルの学習時には、学習用の大量の文書データを文書記憶部１０１に記憶させておき、文書記憶部１０１から読み出した文書データは記事、段落、一文等の適当な単位ごとに文書解析部１０２に読み込まれ、文書解析部１０２でその文書データを解析して単語が抽出される。抽出された単語に関連している概念識別子を特徴ベクトル生成用辞書１０９を参照して求め、ここで求められた概念識別子の列をもとにして概念ベクトル生成部１０３で概念識別子の特徴ベクトルを生成し、１０３で生成された概念識別子の特徴ベクトルは概念ベクトル記憶部１０４に記憶される。こうして概念識別子の特徴ベクトルを学習する。

【００２５】文書の分類をする時には、分類する文書のデータを文書記憶部１０１に記憶させておき、文書記憶部１０１から読み出した文書データは分類を行なわせたい単位（例えば記事単位）ごとに文書解析部１０２に読み込まれ、文書解析部１０２でその文書データの解析をして単語が抽出される。ここで抽出された単語に関連している概念識別子の特徴ベクトルを概念ベクトル記憶部１０４の内容を参照して求める。通常は文書データの一つの単位（例えば一つの記事）から複数の単語が抽出され、それに関連する概念識別子も複数になるが、この場合には関連するすべての概念識別子の特徴ベクトルの値を平均化することで文書の特徴ベクトルが計算される。

【００２６】この時、単純に平均化するのではなく、各概念識別子の特徴ベクトルをその概念識別子の出現頻度の逆数に応じて重み付けをしてから（例えば、大量の記事からその概念識別子の出現している記事数を調査し、ｌｏｇ（全記事数／その概念識別子の出現している記事数）をその概念識別子の特徴ベクトルに乗じてから）平均化するとより良い値が得られる場合がある。

【００２７】文書の特徴ベクトルが求まったら従来のクラスタリングの手法を適用することで文書の分類を行なうことができる。これは例えば文書の特徴ベクトル間の距離が近い文書同士は同じ分野に属するとみなせば良い。

【００２８】また、人間が各分類群ごとに典型的な文書を選び、その文書から抽出される概念識別子の特徴ベクトルからその分類群の仮の代表ベクトルを生成しておき、文書記憶部１０１から読み込まれる文書の特徴ベクトルがどの分類群の仮の代表ベクトルに近いかで文書を分類することもできる。このような分類手法でも１０１から大量に文書データを読み込ませれば仮の代表ベクトルを人間が選んでいるということに起因する誤差の影響が少なくなり、最終的には各分野毎のかなり一般的な代表ベクトルを生成することができる。

【００２９】以下に、概念識別子の特徴ベクトルの生成法を説明する。

【００３０】文書を形態素解析する単語を$word_1$，$word_2$，・・・，$word_p$のp個とし、特徴ベクトルの各要素に対応する概念識別子を$conc_1$，$conc_2$，・・・，$conc_n$のn個とし、特徴ベクトルを持たせる概念識別子（特徴ベクトル作成概念識別子と呼ぶことにする）をcid_1，cid_2，・・・，cid_qのq個とし、概念識別子の特徴ベクトルを学習するために用意された記事はm個あるとする。

【００３１】ここで単語とは、「私」「Ｉ」「ｉｃｈ」など、扱いたい言語の単語であり、概念識別子とは、各概念に付けられた番号である。辞書によっては各単語と関連している概念識別子のリストが得られるようになっているものがある。例えば、（株）日本電子化辞書研究所のＥＤＲ電子化辞書等である。このような辞書が利用できない場合でも、辞書に登録されている単語を、例えばコードの小さい順にソートして各単語に番号を割り振り、「その単語の番号」＝「その単語に関連している概念識別子」とすることで、本分類手法を用いることができる。ここで、各単語に番号を割り振るには、ソートした場合に何行目にくるかをその単語の番号にすれば良い。

【００３２】また、日常的に使用される国語辞典、英和辞典、独和辞典等を利用することによっても各単語と関連している概念識別子を得ることができる。例えば、概念「私」の番号を『私』で表すとすると、単語「私」に関連している概念識別子は『私』となる。ここで、概念「私」の番号は、単語「私」の番号とするなど適当に定めてしまって構わない。英和辞典に『Ｉ：私』という項目があれば、単語「Ｉ」に関連している概念識別子も『私』とすることができる。また独和辞典に「ｉｃｈ：私」という項目があれば、単語「ｉｃｈ」に関連している概念識別子も『私』とすることができる。一般には一つの単語には複数の概念が関連していることがあるので、各単語に関連している概念識別子は複数存在しても良い。

【００３３】記事ｉに含まれる単語の出現頻度ベクトルV_iを

【００３４】

【数１】

$$\mathbf{V}_i = (v_{i1}, v_{i2}, \cdots, v_{ip}) \qquad (1)$$

v_{ij} ： 記事i中に出現する単語$word_j$の個数

【００３５】で表し、記事ｉに含まれる特徴ベクトル作成概念識別子の出現頻度ベクトルU_iを

【００３６】　　【数２】

$$\mathbf{U}_i = (u_{i1}, u_{i2}, \cdots, u_{iq}) \tag{2}$$

u_{ij} ： 記事 i 中に出現する特徴ベクトル作成概念識別子 cid_j の個数

【００３７】で表す。

【００３８】単語ｗｏｒｄｉ と概念識別子ｃｏｎｃｊ との関連の強さを返す関数をｆ（ｗｏｒｄｉ，ｃｏｎｃｊ）とする。使用する概念辞書によっては関連の強さが記述されていない場合があるが、この場合は単語ｗｏｒｄｉ と概念識別子ｃｏｎｃｊ とが関連していればｆ（ｗｏｒｄｉ，ｃｏｎｃｊ）＝１、単語ｗｏｒｄｉ と概念識別子ｃｏｎｃｊ とが関連していなければｆ（ｗｏｒｄｉ，ｃｏｎｃｊ）＝０と定義する。

【００３９】一つの単語には複数の概念識別子が関連付けられている場合があるが、概念識別子の出現頻度ベクトルＵｉ を作成する時に、これを全部使う方法と、一つあるいは適当な個数まで使う方法とがある。つまり、より一般的には複数の概念識別子のうちｒ個までを使うということにすれば、これらの全ての場合に対応できる。例えば、全ての概念識別子を使いたければｒ＝ｎにすれば良いし、一つだけ使いたければｒ＝１とすれば良い。そこで、記事ｉに含まれている単語に関連付けられている概念識別子のうちｒ個までを扱う場合の概念識別子出現頻度ベクトルをＴｉ で表すことにすると次のように定義される。

【００４０】

【数３】

$$\mathbf{T}_i = (t_{i1}, t_{i2}, \cdots, t_{in}) \tag{3}$$

$$t_{ij} = \sum_{k=1}^{p} v_{ik} \cdot g(word_k, conc_j)$$

$$g(word_i, conc_j) = \begin{cases} f(word_i, conc_j), & \text{for } h(word_i, conc_j) \in \mathrm{rmax}_{k=1}^{n,r} h(word_i, conc_k) \\ 0, & \text{for } h(word_i, conc_j) \notin \mathrm{rmax}_{k=1}^{n,r} h(word_i, conc_k) \end{cases}$$

ただし $h(word_i, conc_j)$ は、全 $f(word_i, conc_j)(1 \le j \le n)$ の中で値が同じものが存在しないように値を微調整したもの。簡単には $h(word_i, conc_j) = f(word_i, conc_j) + j/L$（$L$ は十分大きな数）で良い。また、$\mathrm{rmax}_{k=1}^{n,r}$ は、k が１から n まで動く時の各値を大きい順に並べ、上位 r 位までに入るものを集めた集合。

【００４１】すると、特徴ベクトル作成概念識別子ｃｉｄｊ の特徴ベクトルＷｊ は、以下の式で表される。　　【００４２】

【数４】

$$\mathbf{W}_j = (w_{j1}, w_{j2}, \cdots, w_{jn}) = \sum_{i=1}^{m} u_{ij} \cdot \frac{\mathbf{T}_i}{|\mathbf{T}_i|} \tag{4}$$

【００４３】この式からわかるように、全記事について概念識別子の出現頻度ベクトルＴｉをその記事中での出現頻度分の重み付きで加算していくため、特徴ベクトル作成概念識別子ｃｉｄｊ の特徴ベクトルＷｊ は特徴ベクトル作成概念識別子ｃｉｄｊが頻繁に含まれる記事の分野の概念識別子出現頻度分布に類似した値を持つことになる。

【００４４】記事の特徴ベクトルＡ１，Ａ２，・・・，Ａｎ は、概念識別子の特徴ベクトルから以下の式で算出される。

【００４５】

【数５】

$$\mathbf{A}_i = \sum_{j=1}^{q} \log\left(\frac{m}{m_j}\right) \cdot u_{ij} \cdot \frac{\mathbf{W}_j}{|\mathbf{W}_j|} \tag{5}$$

m_j ： 特徴ベクトル作成概念識別子 cid_j が含まれている記事の個数

【００４６】なお、特徴ベクトルを持たせる概念識別子と、特徴ベクトルの各要素の対応する概念識別子とは全く同一のものにしても良いし、全く別のものにしても良い。例えばベクトルの次元数は１００程度にして、特徴ベクトルを持たせる概念識別子を１０００程度にすることもできる。以下の具体的な説明の際にはわかりやすくするために、全く同一のものを使用する。つまり、ｎ＝ｑであり、すべてのｉ（１＜ｉ≦ｎ）において、ｃｏｎｃｉ ＝ｃｉｄｉ であるとする。

【００４７】以下に、具体的に概念識別子の特徴ベクトルの生成法を説明する。

【００４８】例文Ａ「アメリカ政府が先進主要国にココ

ム規制の抜本的な見直しを提案してきた。」
例文Ｂ「規制対象国が兵器の製造につながる工業製品の輸出を規制することを条件に、ココムの規制品目を大幅に削減する意向のようだ。」
という文書データからどのように概念識別子の特徴ベクトルを作成するかを説明する。ここでは、文書データは「一文」という単位で読み込まれることとするが、これは一記事など他の単位でも構わない。
【００４９】また、特徴ベクトルの次元数が２１次元、すなわち、特徴ベクトル生成用辞書に登録されている概念識別子の個数が２１個で、各要素は『アメリカ』『政府』『進んでいる』『おもな』『国家』『ココム』『規制』『抜本的』『見直し』『提案』『対象』『戦いの道具』『作り出す』『工業』『製品』『輸出』『条件』『品物の種類』『大幅』『削減』『意向』という概念識別子に対応しているとする。（『言葉』は、「言葉」という概念に付けられている概念識別子を表す。実際に使用する時にはなんらかの数値であるが、特にここで数値を限定する必要はないので、このように自由度の高い形で記述している。）そして、「アメリカ」「政府」「先進」「主要」「国」「ココム」「規制」「抜本的」「見直し」「提案」「対象」「兵器」「製造」「工業」「製品」「輸出」「条件」「品目」「大幅」「削減」「意向」という各単語に関連付けられている概念識別子は以下のようになっているとする。
【００５０】単語「アメリカ」に対して概念識別子『アメリカ』
単語「政府」に対して概念識別子『政府』
単語「先進」に対して概念識別子『進んでいる』
単語「主要」に対して概念識別子『おもな』
単語「国」に対して概念識別子『国家』
単語「ココム」に対して概念識別子『ココム』
単語「規制」に対して概念識別子『規制』
単語「抜本的」に対して概念識別子『抜本的』
単語「見直し」に対して概念識別子『見直し』
単語「提案」に対して概念識別子『提案』
単語「対象」に対して概念識別子『対象』
単語「兵器」に対して概念識別子『戦いの道具』
単語「製造」に対して概念識別子『作り出す』
単語「工業」に対して概念識別子『工業』
単語「製品」に対して概念識別子『製品』
単語「輸出」に対して概念識別子『輸出』
単語「条件」に対して概念識別子『条件』
単語「品目」に対して概念識別子『品物の種類』
単語「大幅」に対して概念識別子『大幅』
単語「削減」に対して概念識別子『削減』
単語「意向」に対して概念識別子『意向』
このような条件のもとで、例文Ａが１０１から読み込まれると、１０２で解析されて「アメリカ」「政府」「先進」「主要」「国」「ココム」「規制」「抜本的」「見直し」「提案」という単語が抽出される。各単語は１０３にてそれぞれ概念識別子『アメリカ』『政府』『進んでいる』『おもな』『国家』『ココム』『規制』『抜本的』『見直し』『提案』に変換され、概念識別子の出現頻度分布ベクトルが求められる。
【００５１】これから得られる概念識別子の出現頻度分布ベクトルは
（１，１，１，１，１，１，１，１，１，１，０，０，０，０，０，０，０，０，０，０，０）＝V_A
である。すると、『アメリカ』『政府』等、例文Ａに出現する概念識別子の特徴ベクトルには（１，１，１，１，１，１，１，１，１，１，０，０，０，０，０，０，０，０，０，０，０）＝V_A を加算する。正確には、このベクトル＝V_Aの絶対値を１に正規化したものを加算する。図４は、例文Ａを読み込んだ後の概念識別子の特徴ベクトルを並べて行列にしたものである。
【００５２】次に例文Ｂが文書記憶部１０１から読み込まれると、文書解析部１０２で解析されて「規制」「対象」「国」「兵器」「製造」「工業」「製品」「輸出」「規制」「条件」「ココム」「規制」「品目」「大幅」「削減」「意向」という単語が抽出される。各単語は概念ベクトル生成部１０３にてそれぞれ概念識別子『規制』『対象』『国家』『戦いの道具』『作り出す』『工業』『製品』『輸出』『規制』『条件』『ココム』『規制』『品物の種類』『大幅』『削減』『意向』に変換される。
【００５３】これから得られる概念識別子の出現頻度分布ベクトルは
（０，０，０，０，１，１，３，０，０，０，１，１，１，１，１，１，１，１，１，１，１）＝V_B
である。『規制』は３回出現しているので、この概念識別子の出現頻度分布ベクトル＝V_B を３倍したベクトルである（０，０，０，０，３，３，９，０，０，０，３，３，３，３，３，３，３，３，３，３，３）を『規制』の特徴ベクトルに加算する。正確には、ベクトルV_B の絶対値を１に正規化したものを３倍したベクトルを加算する。『対象』『国家』等、例文Ｂに１回しか出現しない概念識別子の特徴ベクトルには（０，０，０，０，１，１，３，０，０，０，１，１，１，１，１，１，１，１，１，１，１）＝V_B を加算する。正確には、このベクトル＝V_B の絶対値を１に正規化したものを加算する。図５は、例文Ａ、Ｂを読み込んだ後の概念識別子の特徴ベクトルを並べて行列にしたものである。である。
【００５４】なお、図ではわかりやすくするため、以上のように常に整数を加算することにして説明したが、この方法では文の長さによって加算するベクトルの大きさが変化してしまうので、実際には、加算するベクトルの絶対値を１に正規化したり、出現頻度分布のベクトルの絶対値を１に正規化してから出現数に比例した値を掛け

た後に加算する方法をとった方が良い。この方法については、これまでの説明の中で、「正確には、」として記述した。

【００５５】そして最終的に得られた特徴ベクトルは、絶対値を１に正規化しておく。

【００５６】こうして得られた概念識別子の特徴ベクトルは概念ベクトル記憶部１０４に記憶され、文書の分類時に利用される。具体例として以下の例文Ｃが読み込まれた時の処理を説明する。

【００５７】例文Ｃ「アメリカ政府は兵器の削減を提案した。」

例文Ｃが１０１から読み込まれると、文書解析部１０２で解析されて「アメリカ」「政府」「兵器」「削減」「提案」という単語が抽出される。各単語は文書ベクトル生成部１０５にてそれぞれ概念識別子『アメリカ』『政府』『戦う道具』『提案』に変換される。

【００５８】すると文書ベクトル生成部１０５では概念ベクトル記憶部１０４の内容を参照して『アメリカ』『政府』等、例文Ｃに出現する概念識別子の特徴ベクトルを加算していき、例文Ｃの特徴ベクトルとして

（３，３，３，３，５，５，９，３，３，３，２，２，２，２，２，２，２，２，２，２）

を得る。図６は、図５に示した概念識別子の特徴ベクトルを利用して例文Ｃの特徴ベクトルを生成した結果を示す。である。図６ではわかりやすさを優先するためにベクトルの正規化を行なっていないが、実際の処理では加算する前に各概念識別子の特徴ベクトルの絶対値を１に正規化してから加算を行ない、最後に得られた特徴ベクトルの絶対値も１に正規化しておく。

【００５９】次に、分類時に文書の特徴ベクトルをどのように利用するのかを説明する。簡単には、まず求まった文書の特徴ベクトルの絶対値を１に正規化してから、Ｋ－ｍｅａｎｓ法などの従来からある手法を用いて分類したり、分類群の（仮）代表ベクトルとの類似度で分類すれば良い。ここで、類似度は、距離を求めたり内積を計算することによって得られる。

【００６０】分類の具体例として、分類群が３つあり、それぞれの分類群の代表ベクトルが以下のように求められていたとする。

【００６１】

【数６】

$$\text{分類群１の代表ベクトル } (1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1)$$
$$\text{分類群２の代表ベクトル } (1,1,1,1,1,1,1,1,1,1,5,5,5,5,5,5,5,5,5,5)$$
$$\text{分類群３の代表ベクトル } (4,4,4,4,6,6,6,3,3,3,1,1,1,1,1,1,1,1,1,1)$$

【００６２】類似度の尺度として、文書の特徴ベクトル、分類群の代表ベクトル共に絶対値を１に正規化してから両者の内積を計算し、一番大きな値をとるものが一番類似度が高いとすると、

【００６３】

【数７】

$$\text{例文Ｃの特徴ベクトル} \frac{1}{\sqrt{234}}(3,3,3,3,5,5,9,3,3,3,2,2,2,2,2,2,2,2,2,2)$$
$$\text{分類群１の代表ベクトル} \frac{1}{\sqrt{6}}(1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1)$$
$$\text{分類群２の代表ベクトル} \frac{1}{\sqrt{285}}(1,1,1,1,1,1,1,1,1,1,5,5,5,5,5,5,5,5,5,5)$$
$$\text{分類群３の代表ベクトル} \frac{1}{\sqrt{210}}(4,4,4,4,6,6,6,3,3,3,1,1,1,1,1,1,1,1,1,1)$$

【００６４】なので、例文Ｃの特徴ベクトルと各分類群の代表ベクトルとの内積は

【００６５】

【数８】

$$\text{分類群１との内積} \frac{1}{\sqrt{235}} \cdot \frac{1}{\sqrt{6}} \cdot 20 \simeq 0.4583$$
$$\text{分類群２との内積} \frac{1}{\sqrt{236}} \cdot \frac{1}{\sqrt{285}} \cdot 150 \simeq 0.5759$$
$$\text{分類群３との内積} \frac{1}{\sqrt{238}} \cdot \frac{1}{\sqrt{210}} \cdot 211 \simeq 0.9438$$

【００６６】となり、例文Ｃの特徴ベクトルは分類群３の代表ベクトルに一番近いことがわかるので、例文Ｃは分類群３に分類される。図７は、この結果を示す。図７も図６と同様わかりやすさを優先するためにベクトルの正規化を行なっていないが、実際の処理では比較を行なう前に各ベクトルの絶対値を１に正規化してから比較を行なう。

【００６７】この分類装置の有効性を評価するための実験を行なった。

【００６８】実験方法として、具体的には、１９８７年の朝日新聞の４００記事を、「政府」「経済」「国際」「社会（犯罪，事件）」「社会（教育，人間）」の五つの分野に分類するタスクにおいて、人間が分類した結果を正解として分類正解率を求めた。但し、二つの分野に分類できるような記事は、そのどちらに分類されても正解とみなした。

【００６９】このタスクを行なう場合の特徴ベクトルの構成時に、

１．単語をそのまま用いる

２．ＥＤＲの辞書を利用して求めた概念識別子を用いるの２種類の方法で分類正解率を比較した。

【００７０】実験に使用したデータとその使用目的等は以下の通りである。

【００７１】１．ＥＤＲ電子化辞書評価版第２．１版

単語データの抽出にＥＤＲ電子化辞書の日本語単語辞書評価版第２．１版を使用し、単語間の類似度の計算に上

記日本語単語辞書とＥＤＲ電子化辞書の概念辞書評価版第２．１版を使用した。

【００７２】評価版第２．１版の日本語単語辞書の登録語数は基本語約１６万語、専門用語約４．２万語であり、評価版第２．１版の概念辞書の収録概念数は約３６万概念である。

【００７３】２．ＣＤ－ＨＩＡＳＫ（朝日新聞のＣＤ－ＲＯＭ）１９９０年版（約１５０Ｍバイト、１０１９６記事）

特徴ベクトルを生成するためのデータとして使用した。また各分野の典型的な記事もここから抜き出し、各分野の代表ベクトルを生成する時にも使用した。

【００７４】３．ＣＤ－ＨＩＡＳＫ（朝日新聞のＣＤ－ＲＯＭ）１９８７年版から抜き出した記事

人間による分類と分類装置による分類との比較に使用した。この記事の一例を以下に示す。

【００７５】物価と為替の安定維持が最大の課題　澄田日銀総裁が語る　澄田日銀総裁は３１日、朝日新聞とのインタビューで、新年の金融政策について、物価の安定維持が最大の課題であることを強調しつつ、内需拡大、対外不均衡の是正に取り組む姿勢を明らかにした。その一方で、日本経済が国際的に影響力を増していることを踏まえ、国際協調がますます重要になっていることを指摘しながらも、金融政策が外圧や国内政治からの独立性と自主性を確保することが一層……

４．単語データ

ＥＤＲ電子化辞書評価版第２．１版の日本語単語辞書中から「平仮名だけからなる三文字以下の単語」と「漢字以外の一文字単語」を除いた全単語を使用した。文書から単語を抽出する方法は、長さの長いものを優先して選択するパターンマッチング（最長一致法）によるが、連続する二単語を複数の組み合わせで抽出できる場合には、その二単語の合計の長さが最長になる組み合わせの最初の単語を選択する手法（二文節最長一致法）を用いた。ただし誤抽出をできるだけ減らすため、漢字一文字の単語の場合は前後が非漢字の場合のみ抽出した。朝日新聞１９９０年１月１日朝刊の最初から抽出された５００単語について調査した結果、この方法で９５％程度正しく抽出されることを確認した。

【００７６】５．概念識別子データ

ＥＤＲ電子化辞書評価版第２．１版の日本語単語辞書中から「平仮名だけからなる三文字以下の単語」と「漢字以外の一文字単語」を除いた全単語について、関連する概念識別子を調査し、使用頻度の高いものを採用した。関連する概念識別子を全部使用する場合と、一つだけ使用する場合との二通りで実験を行なった。

【００７７】６．特徴ベクトル

特徴ベクトルの各要素に対応する概念識別子（または単語）と、特徴ベクトルが付加される概念識別子（または単語）とは、同一のものをしようすることにした。

【００７８】また、特徴ベクトルの次元数は４０９６，２０４８，１０２４，５１２，２５６，１２８，６４の７種類で実験した。

【００７９】この次元数個分の概念識別子（または単語）の選出の方法であるが、単純に朝日新聞１９９０年版の中で出現頻度の高いものから順番に選出した。

【００８０】このようにして得られた、実験結果について以下に説明する。

【００８１】分類する時に、分野の第１位候補と第２位候補とのスコア（記事の特徴ベクトルと、分野の代表ベクトルとの内積の値）の比の大小によって、分野がまぎらわしいかはっきりしているかを判定できるので、スコアの比が大きいもの（分野がはっきりしているもの）は分類の易しい記事群、スコアの比が小さいもの（分野がまぎらわしいもの）は分類の難しい記事群として、分類の正解率は分類の易しい記事群（２００記事）、難しい記事群（２００記事）、全体（４００記事）、の３つの値を出した。この結果を表１に示す。

【００８２】

【表１】

表 1: 分類正解率 [%]

次元数		4096	2048	1024	512	256	128	64
単語を	易	98.0	98.0	96.0	96.5	92.0	84.0	80.5
そのまま	難	67.5	69.0	66.0	57.5	57.5	48.5	44.0
利用	全体	82.75	83.5	81.0	77.0	74.75	65.75	62.25
概念識別	易	98.5	98.5	97.0	94.0	89.0	82.0	76.5
子を全部	難	67.5	67.5	62.5	57.0	42.5	43.0	43.0
利用	全体	83.0	83.0	79.75	75.5	65.75	62.5	59.75
概念識別	易	98.5	99.0	97.0	95.5	88.5	84.0	82.5
子を一つ	難	70.5	70.0	66.5	63.5	43.5	43.0	47.5
だけ利用	全体	84.5	84.5	81.75	79.5	66.0	63.5	65.0

【００８３】表１より、ベクトルの次元数が５１２以上の時には概念識別子を一つだけ用いると、かなり高い精度での分類ができることが確認できる。特に次元数２０４８における易しい記事の分類正解率は９９％となり、ほぼ１００％近い正解率で分類できることがわかる。これはＥＤＲ電子化辞書評価版第２．１版をそのまま使用した場合の実験結果だが、他の辞書を用いれば辞書の単語や概念識別子の粒度に応じて最適なベクトルの次元数

が変化することが予想される。

【００８４】つまり、ベクトルの次元数を高くとれる場合には、粒度の細かい辞書を使用し、ベクトルの次元数があまり高くとれない場合には、粒度の粗い辞書を使用すると、高い正解率が得られることが予想されるため、分類装置が使用できる記憶容量に応じて、その容量にあった粒度の辞書を使用すると良い。

【００８５】請求項２に記載の発明の文書分類装置の一実施例を図２に示す。ここで、図２（ａ）は、全体の装置構成、図２（ｂ）は、学習時に使用される装置の構成、図２（ｃ）は、分類時に使用される装置の構成を夫々示す。図中、２０１は文書記憶部、２０２は文書解析部、２０３は概念ベクトル生成部、２０４は概念ベクトル記憶部、２０５は文書ベクトル生成部、２０６は文書ベクトル記憶部、２０７は分類部、２０８は結果記憶部、２０９は特徴ベクトル生成用辞書、２１０は有用概念識別子選出部である。

【００８６】図１に示した実施例と同様の方法によって、概念識別子の特徴ベクトルを学習し、それをもとに大量の文書データを分類する。分類した結果は結果記憶部２０８に記憶されているが、この結果を元にして、有用概念識別子選出部２１０で有用概念識別子の選出を行なう。これは、分類群ごとに各概念識別子の出現頻度を求め、どの分類群にも同じような割合で含まれている概念識別子を除去したり（方法１：最高頻度と最低頻度との比がある閾値以下のものを除去）、ある分類にだけ高い割合で含まれているものを選出したり（方法２：最高頻度と第二位頻度との比がある閾値以上のものを選出）する。なお、有用概念識別子選出部２１０で選出を行なう概念識別子は必ずしも特徴ベクトル生成用辞書２０９に登録されている概念識別子からでなくても良く、もっと広い範囲の概念識別子から選出を行なうことができる。

【００８７】具体例として分類群がａ，ｂ，ｃの三つあったとして、特徴ベクトル生成用辞書２０９に登録されている概念識別子が『政治』『日本』『国際』の三つだったとする。そして分類群ごとに各概念識別子（特徴ベクトル生成用辞書２０９に登録されている概念識別子以外に『選挙』『問題』についても頻度を調べるとする）の頻度が次のようだったとする。

【００８８】分類群ａ　政治３０％，日本５％，国際３５％，選挙１０％，問題２０％

分類群ｂ　政治３％，日本５５％，国際３５％，選挙２％，問題５％

分類群ｃ　政治３％，日本３０％，国際３５％，選挙２％，問題３０％

この場合に、方法１を用いると『国際』はどの分類群にも同じような割合で含まれているので、特徴ベクトル生成用辞書から除去することになる。『政治』『日本』『選挙』『問題』は分類群ごとの頻度に偏りがあるので、有用概念識別子として選出され、特徴ベクトル生成用辞書２０９に登録する（この時登録概念識別子数を抑えたい場合は、頻度に偏りのある概念識別子の中で、合計の出現頻度の順番に登録したい個数だけ取ってくれば良い。）方法２を用いた場合『政治』と『選挙』だけが選出され特徴ベクトル生成用辞書２０９に登録し、『日本』や『国際』や『問題』は特徴ベクトル生成用辞書２０９には登録しない。方法１と方法２の中間的な方法として、第１位の頻度と第ｎ位（ｎは３以上、分類群の個数－１以下）の頻度との比がある閾値以上であるかどうかで有用概念識別子を選出する方法も考えられる。

【００８９】また、頻度の比ではなく、頻度の分散の値が大きいものを選出する方法も考えられる。

【００９０】なお、このようにして選出された概念識別子は頻度の比（あるいは頻度の分散）に応じた重要度を持っていると考えることができるので、文書の特徴ベクトルを計算する時にはその文書内の概念識別子の特徴ベクトルをこの比（あるいは分散）に応じて重み付けをしてから（例えばｌｏｇ（頻度の比）をその概念識別子の特徴ベクトルに掛けてから）平均化するとより良い文書の特徴ベクトルの値が得られる場合がある。

【００９１】こうして特徴ベクトル生成用辞書２０９に、分類に有用な概念識別子だけを登録し、もう一度、概念識別子の特徴ベクトルを学習し、それを用いて文書を分類すると、特徴ベクトル生成辞書をより小さくできたり、分類の精度をあげることができる。

【００９２】請求項３に記載した発明の文書分類装置の一実施例を図３に示す。ここで、図３（ａ）は、全体の装置構成、図３（ｂ）は、学習時に使用される装置の構成、図３（ｃ）は、分類時に使用される装置の構成を夫々示す。図中３０１は文書記憶部、３０２は文書解析部、３０３は概念ベクトル生成部、３０４は概念ベクトル記憶部、３０５は文書ベクトル生成部、３０６は文書ベクトル記憶部、３０７は分類部、３０８は結果記憶部、３０９は特徴ベクトル生成用辞書、３１０は有用概念識別子選出部、３１１は代表ベクトル生成部、３１２は代表ベクトル記憶部である。図１に示した実施例を基にして、本実施例の装置を構成する場合には有用概念識別子選出部３１０が無いシステムとなる。

【００９３】図１及び図２に示した実施例と同様の方法によって、概念識別子の特徴ベクトルを学習し、それをもとに大量の文書データを分類する。分類した結果は３０８に記憶されているが、この結果を元にして、３１１で代表ベクトルを生成する。これは、分類群ごとの各概念識別子の頻度を求め、各概念識別子の特徴ベクトルを頻度の重みをつけて平均したものである。具体例として分類群がａ，ｂ，ｃの三つあったとして、特徴ベクトル生成用辞書３０９に登録されている概念識別子が『政治』『国会』『国際』の三つだったとする。そして分類群ごとの各概念識別子の頻度が次のようだったとする。

【００９４】
分類群ａ　政治４０％，国会５０％，国際１０％
分類群ｂ　政治１０％，国会１０％，国際８０％
分類群ｃ　政治２０％，国会１０％，国際７０％
すると、分類群ａの代表ベクトルは、『政治』の特徴ベクトルに０．４を掛けたものと、『国会』の特徴ベクトルに０．５を掛けたものと、『国際』の特徴ベクトルに０．１を掛けたものの和として与えられる。
【００９５】また、分類群ａに分類された文書全部の特徴ベクトルの平均をとったものを分類群ａの代表ベクトルとする方法も考えられる。
【００９６】こうして、代表ベクトルが生成されたらそれを代表ベクトル記憶部３１２に記憶しておくことで、以後の文書の分類時にはこの代表ベクトルを参照することで、文書記憶部３０１から読み込まれた文書は、その文書の特徴ベクトルにもっとも類似した代表ベクトルに対応する分類群に分類することができるようになる。これにより、分類の処理が高速化できる。
【００９７】本実施例を用いて文書を分類している様子の一例を図８に示す。これは、分類装置側の「分類したい文を入力して下さい。」という質問に対して、ユーザが「大手保険会社の債券投資姿勢に格差が生じてきた。」という文を入力した場合の例である。このユーザ入力文の特徴ベクトルと分類群「政治」の代表ベクトルとの類似度は約０．４５８３、分類群「国際」の代表ベクトルとの類似度は約０．５７５９、分類群「経済」の代表ベクトルとの類似度は約０．９４３８となり、このユーザ入力文はもっとも類似度の高い分類群「経済」に分類されている。
【００９８】請求項５に記載した文書検索装置の一実施例を図１０に示す。ここで、図１０（ａ）は、全体の装置構成、図１０（ｂ）は、学習時に使用される装置の構成、図１０（ｃ）は、検索時に使用される装置の構成を夫々示す。図中１００１は文書記憶部、１００２は文書解析部、１００３は概念ベクトル生成部、１００４は概念ベクトル記憶部、１００５は文書ベクトル生成部、１００６は文書ベクトル記憶部、１００７は検索部、１００８は出力部、１００９は特徴ベクトル生成用辞書、１０１０は検索文入力部である。
【００９９】文書記憶部１００１には、学習に用いるための文書や、検索対象の文書を記憶する。検索文入力部１０１０からは、検索したい文（単語だけでも良い）が入力される。文書解析部１００２は文書記憶部１００１や検索文入力部１０１０から文書を渡され、特徴ベクトル生成用辞書１００９中の単語辞書を用いてその文書の形態素解析（単語等に分けること）を行なう。
【０１００】概念ベクトルを学習する時の各構成要素の作用の概要を、図１０（ｂ）に基づいて説明する。概念ベクトル生成部１００３では、文書解析部１００２から渡された単語データを、特徴ベクトル生成用辞書１００９中の概念辞書（単語と概念識別子との関連付けを行なっている辞書）を参照して概念識別子に変換し、概念識別子間の共起関係を利用して概念識別子の特徴ベクトルを生成する。概念ベクトル記憶部１００４は、概念ベクトル生成部１００３で生成された概念識別子の特徴ベクトルを記憶する。
【０１０１】文書を検索する時の各構成要素の作用の概要を、図１０（ｃ）に基づいて説明する。文書ベクトル生成部１００５では、文書解析部１００２から渡された単語データを、特徴ベクトル生成用辞書１００９中の概念辞書を参照して概念識別子に変換し、そこで得られた概念識別子の特徴ベクトルを概念ベクトル記憶部１００４を参照して求め、文書中から得られる全ての単語についてこのようにして求めた概念識別子の特徴ベクトルから、平均化するなどして文書の特徴ベクトルを求める。文書ベクトル記憶部１００６は、文書ベクトル生成部で求められた文書の特徴ベクトルを記憶する。検索文入力部１０１０から入力された文も、同様にして特徴ベクトルが求められ、文書ベクトル記憶部１００６には、検索文の特徴ベクトルも記憶される。検索部１００７は、文書ベクトル記憶部１００６から検索文の特徴ベクトルを取得し、文書ベクトル記憶部１００６に記憶されている各文書の特徴ベクトルと検索文の特徴ベクトルとの類似度が高いものを検索結果として出力部１００８に渡す。出力部１００８では、検索部１００７から渡された検索結果を出力する。
【０１０２】検索部１００７での類似度の高さの判定は、検索文の特徴ベクトルの絶対値を１に正規化したものと、各文書の特徴ベクトルの絶対値を１に正規化したものとの内積をとって判断する。内積がある閾値（例えば０．９）より高いものを検索結果として出力部１００８に渡す方法や、内積の高い順番に文書を適当な個数（例えば１０個）選出して出力部１００８に渡す方法等がある。
【０１０３】この実施例での曖昧検索の一例を図１１に示す。ここで、図１１（ａ）は日本語による曖昧検索の例を示し、図１１（ｂ）は英語による曖昧検索の例を示す。この例では大量の電子メールの中から、検索文「歌を歌いたい」と検索文「Ｉｗａｎｔ　ｔｏ　ｓｉｎｇ」とで、どちらもカラオケ関連のメールが検索されることを示している。
【０１０４】同じく、この実施例でのもう一つの曖昧検索例を図１２に示す。この例では、検索文「歌を歌いたい」を入力すると（図１２（ａ））、日本語のメール（カラオケ、図１２（ｂ））と英語のメール（コーラスパーティ、図１２（ｃ））とが検索されることを示している。
【０１０５】請求項４に記載の発明の文書分類装置及び請求項６に記載の発明の文書分類装置に使用される「言語毎の特徴ベクトル生成用辞書」は、各言語毎の単語辞

書と、各言語で共通に用いる概念識別子と各言語の単語との関連を表す概念辞書とを使用したい言語の種類数だけ備える。図９は、複数の言語に対応した特徴ベクトル生成用辞書の概念図を示す。図９では、日本語、英語、ドイツ語という３か国語に対応した特徴ベクトル生成用辞書の例である。例えば、日本語の「私」という単語と、英語の「Ｉ」という単語と、ドイツ後の「ｉｃｈ」という単語が、ともに概念識別子「０００１」　と関連付けられていることを示している。他の単語についても同様である。ただし、この図では「０００１」は「私」という概念に付けられた概念識別子であり、「０００２」は「貴方」という概念に付けられた概念識別子であり、「０００５」は「我々」という概念に付けられた概念識別子であり、「０００６」は「貴方達」という概念に付けられた概念識別子であり、「０１０５」は「赤い」という概念に付けられた概念識別子である。なお、この概念識別子の数値自体は、同じ概念に同じ番号が割り当てられ、違う概念には違う番号が割り当てられていればどんな数値を使っても良いので、本実施例中では「０００１」という直接の数値の代わりに『私』という形で概念識別子を表している。この特徴ベクトル生成用辞書により、入力文書や検索文の言語の種類に応じて辞書を切替えることで、どの言語を用いても共通の概念識別子を用いて分類や検索を行なうことができる。

【０１０６】本発明の文書分類装置及び文書検索装置は、通常の文書の分類や通常の文書の検索にのみ用いられるものではない。すなわち、電子メールや電子ニュースを自動的に分類したり、電子メールや電子ニュースの中からユーザーの興味を持ちそうなものを選出（検索）したり（ユーザーがそれまでに読んだメールやニュースの特徴ベクトルとの類似度で判定できる）、仮名漢字変換における同音異義語の選択（それまでに変換した内容から得られる特徴ベクトルとの類似度で同音異義語を選択する）に利用できる。また、音声認識・手書き文字認識などにおいて過去の文脈に最も適合した変換結果を選択する方法をとる（それまでに認識した内容から得られる特徴ベクトルとの類似度で認識結果を選択する）際や、認識時等において単語等の検索空間を狭める（それまでに認識した内容から得られる特徴ベクトルの平均値に近い概念識別子と関連付けられている単語だけを検索するようにする）際にも利用できる。この場合には、文書記憶部又は、検索入力部に、通常の文書データの代わりに、上記のデータを入力する。また、複数の言語について単語と概念識別子との関連を表す情報があれば、言語の種類を問わずに分類や検索等を行なうことができる。

【０１０７】

【発明の効果】請求項１に記載の文書分類装置によれば、概念識別子の特徴ベクトル及びそれから生成された文書の特徴ベクトルを使用して、文書の学習と学習に基づいた文書の分類が行われる。したがって、文書データを用意するだけで、概念識別子の特徴ベクトルを生成でき、人手を全く必要としない文書自動分類を実現できる。また、概念識別子を用いて特徴ベクトルを生成することで、単純に単語を用いる場合に比べて分類の精度を高めることができる。

【０１０８】請求項２に記載の文書分類装置によれば、分類に有用な概念識別子を用いることによって、特徴ベクトルの記憶空間を削減したり、分類の精度を向上させることができる。

【０１０９】請求項３に記載の文書分類装置によれば、結果記憶部に記憶された分類ごとに、概念識別子や文書の特徴ベクトルを用いて、その分類を代表する文書の特徴ベクトルを求める代表ベクトル生成部と、分類を代表する文書の特徴ベクトルを記憶する代表ベクトル記憶部とをさらに含むように構成されているので、一度各分類群の代表ベクトルを生成してしまえば、新たな文書データを分類するときには、その文書の特徴ベクトルと各分類群の代表ベクトルとの比較を行なうだけでその文書がどの分類群に属すかを判定できるようになる。したがって、分類処理を単純化・高速化できる。

【０１１０】請求項４に記載の文書分類装置によれば、特徴ベクトル生成用辞書が複数の言語の辞書を含んでおり、複数の言語のどの言語の単語であっても同じ概念の単語は同じ概念識別子に変換し、言語の種類によらない文書分類を行うことができる。また、特徴ベクトルは概念識別子に対して生成されるので、言語毎に単語に対して特徴ベクトルを生成する場合に比較して特徴ベクトルの記憶領域を小さく抑えることができる。

【０１１１】請求項５に記載の文書検索装置によれば、概念識別子の特徴ベクトル及びそれから生成された文書の特徴ベクトルを使用して、文書の学習と学習に基づいた文書の検索が行われる。したがって、特徴ベクトルの類似度で文書を検索することで、文字列のパターンマッチングによる検索とは違い、文字列が一致していなくても意味的に類似度が高いものを検索（曖昧検索）することができる。

【０１１２】請求項６に記載の文書検索装置によれば、特徴ベクトル生成用辞書が複数の言語の辞書を含んでおり、複数の言語のどの言語の単語であっても同じ概念の単語は同じ概念識別子に変換し、言語の種類によらない文書検索を行うことができる。

【０１１３】また、特徴ベクトルは概念識別子に対して生成されるので、言語毎に単語に対して特徴ベクトルを生成する場合に比較して特徴ベクトルの記憶領域を小さく抑えることができる。

【０１１４】また、本発明の装置で作成される概念識別子の特徴ベクトルは、文書の分類時や検索時に使えるだけでなく、仮名漢字変換における同音異義語の選択にも利用できるし、音声認識・手書き文字認識などにおい

て、過去の文脈に最も適合した認識結果を選択する方法をとる際にも利用できる。

【図面の簡単な説明】

【図１】本発明の請求項１の基本構成を示すブロック図である。

【図２】本発明の請求項２の基本構成を示すブロック図である。

【図３】本発明の請求項３の基本構成を示すブロック図である。

【図４】本発明の概念識別子の特徴ベクトルの生成を説明する図１である。

【図５】本発明の概念識別子の特徴ベクトルの生成を説明する図２である。

【図６】本発明の文書の特徴ベクトルの生成を説明する図である。

【図７】本発明による文書の分類を説明する図である。

【図８】本発明による文書の分類の例を説明する図である。

【図９】本発明の請求項５の言語毎の複数の特徴ベクトル生成用辞書を説明する図である。

【図１０】本発明の請求項４の基本構成を示すブロック図である。

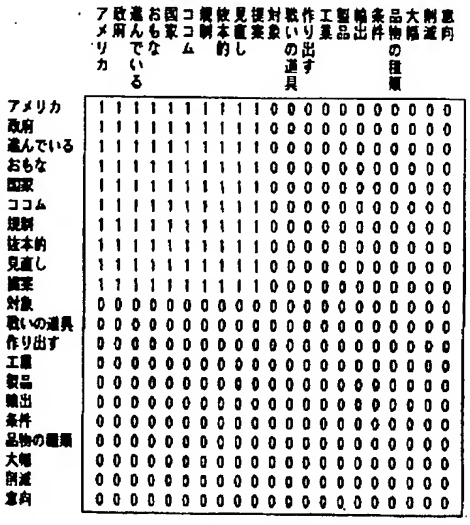
【図１１】本発明の請求項４の実施例による文書検索装置での曖昧検索例を説明する図である。

【図１２】本発明の請求項４の実施例による文書検索装置での曖昧検索例を説明する図２である。

【符号の説明】

１０１　文書記憶部
１０２　文書解析部
１０３　概念ベクトル生成部
１０４　概念ベクトル記憶部
１０５　文書ベクトル生成部
１０６　文書ベクトル記憶部
１０７　分類部
１０８　結果記憶部
１０９　特徴ベクトル生成用辞書

【図４】

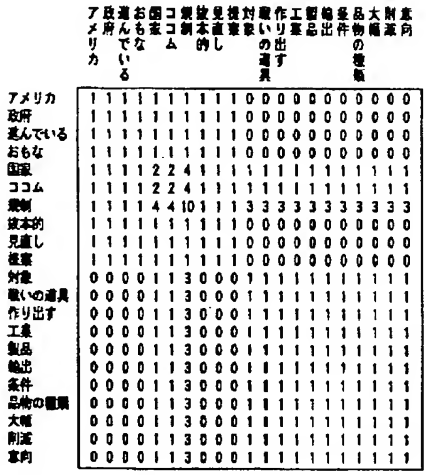


例文Ａ　「アメリカ政府が先進主要国にココム規制の基本的な見直しを提案してきた。」

例文Ａから抽出される概念識別子

『アメリカ』『政府』『進んでいる』『おもな』『国家』『ココム』『規制』『基本的』『見直し』『提案』

【図５】



例文Ａ　「アメリカ政府が先進主要国にココム規制の基本的な見直しを提案してきた。」

例文Ｂ　「規制対象国が兵器の製造につながる工業製品の輸出を規制することを条件に、ココムの規制品目を大幅に削減する意向のようだ。」

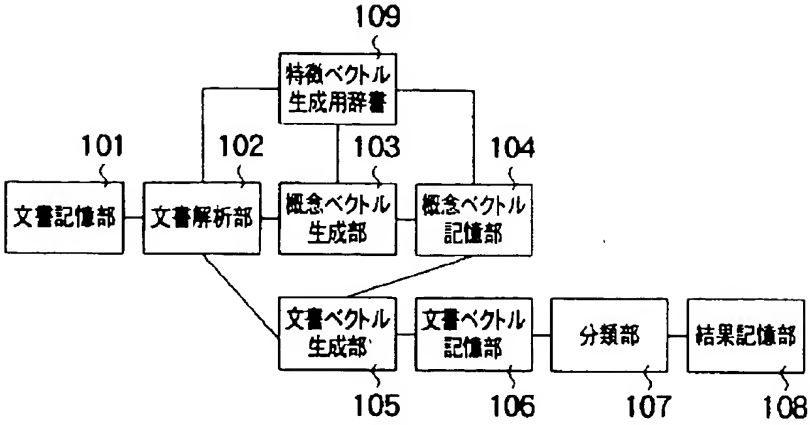
例文Ａから抽出される概念識別子

『アメリカ』『政府』『進んでいる』『おもな』『国家』『ココム』『規制』『基本的』『見直し』『提案』

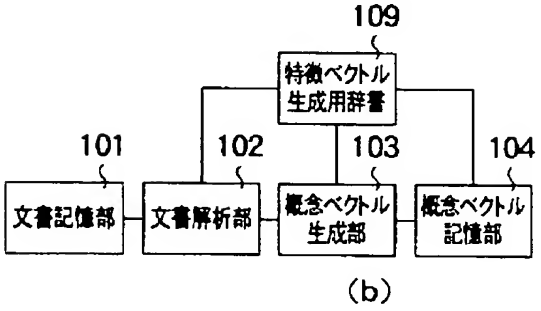
例文Ｂから抽出される概念識別子

『規制』『対象』『国家』『取いの道具』『作り出す』『工業』『製品』『輸出』『規制』『条件』『ココム』『規制』『品物の種類』『大幅』『削減』『意向』

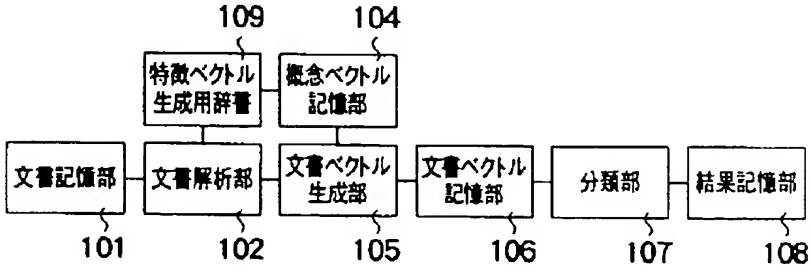
【図1】


109
特徴ベクトル生成用辞書
101
102
103
104
文書記憶部
文書解析部
概念ベクトル生成部
概念ベクトル記憶部
文書ベクトル生成部
文書ベクトル記憶部
分類部
結果記憶部
105
106
107
108

(a)

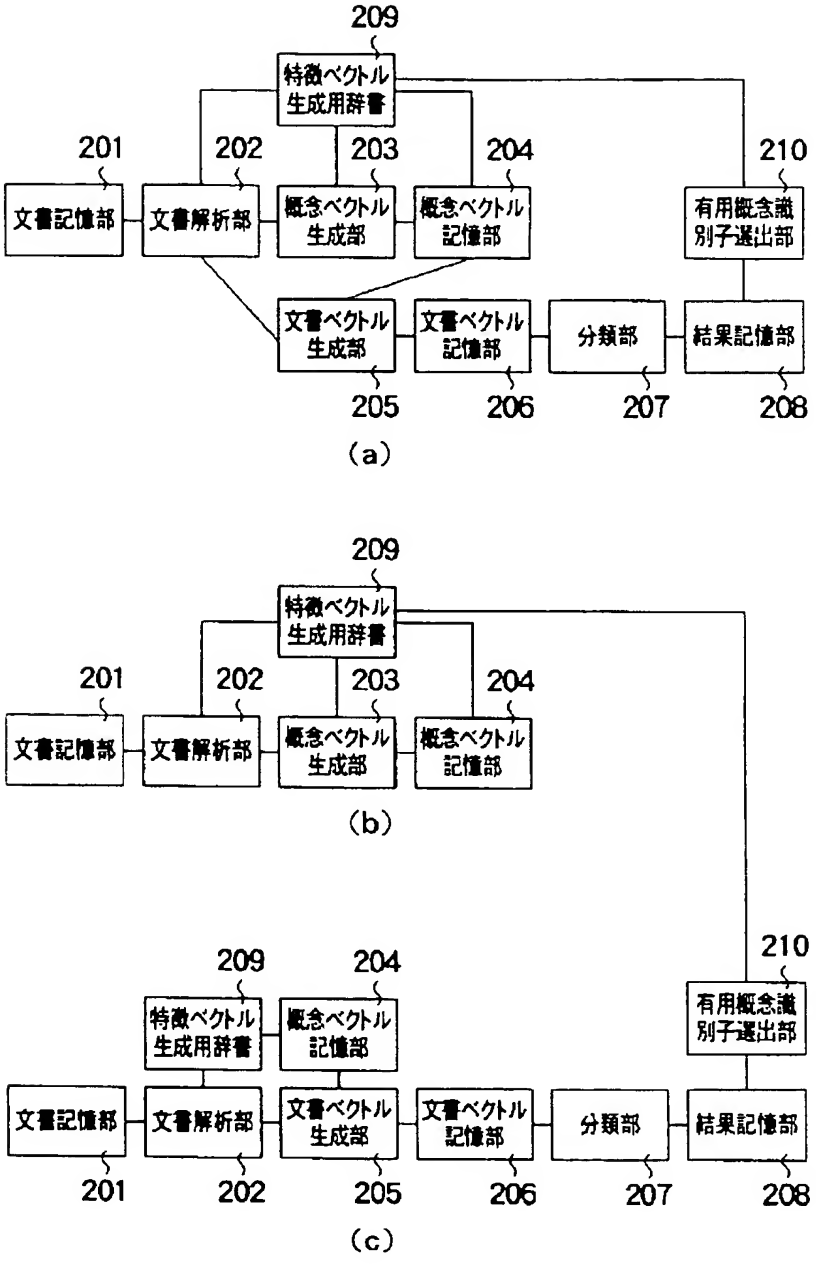

109
特徴ベクトル生成用辞書
101
102
103
104
文書記憶部
文書解析部
概念ベクトル生成部
概念ベクトル記憶部

(b)

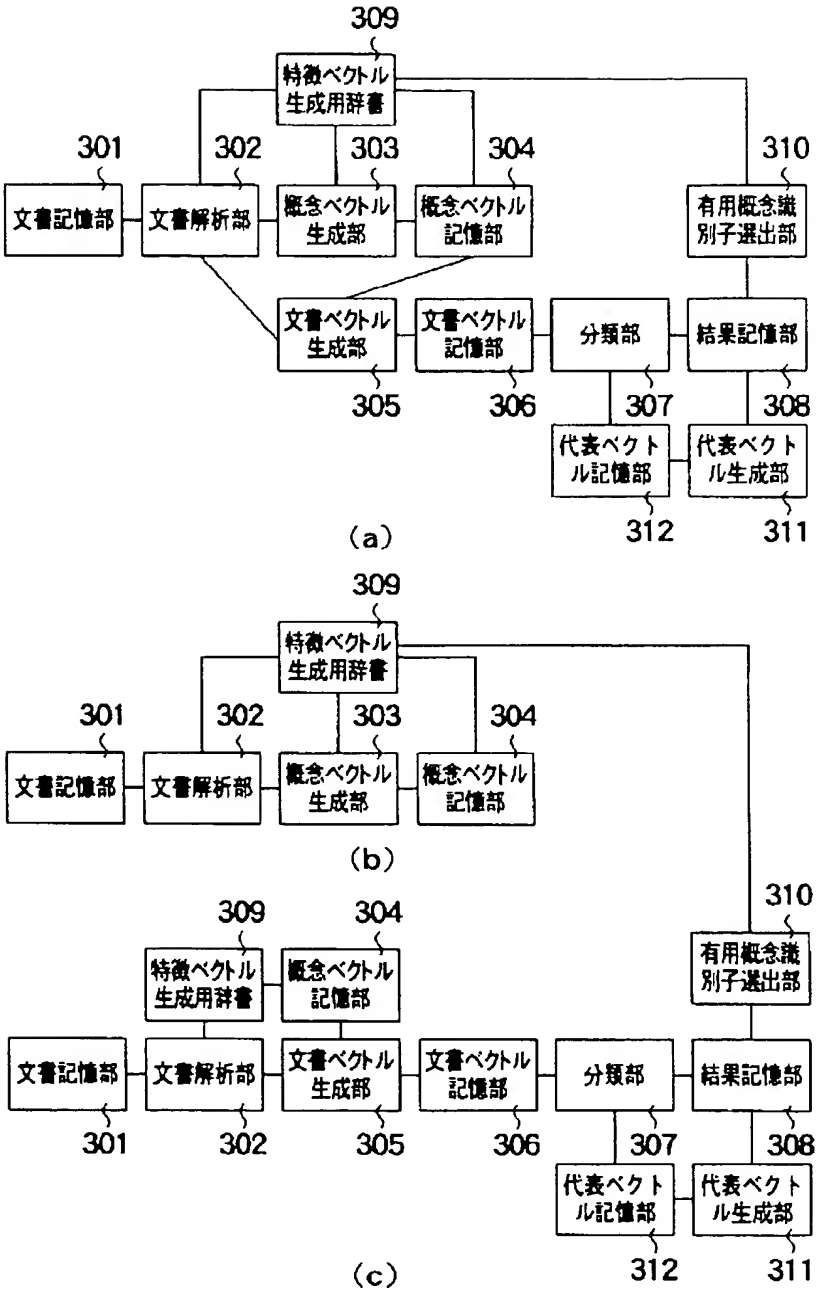

109
104
特徴ベクトル生成用辞書
概念ベクトル記憶部
文書記憶部
文書解析部
文書ベクトル生成部
文書ベクトル記憶部
分類部
結果記憶部
101
102
105
106
107
108

(c)

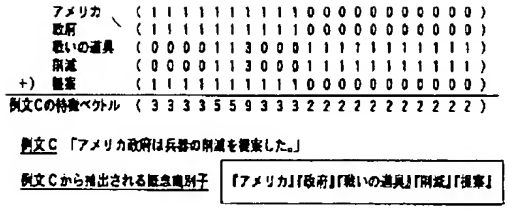
【図2】


209
特徴ベクトル生成用辞書
201
202
203
204
210
文書記憶部
文書解析部
概念ベクトル生成部
概念ベクトル記憶部
有用概念識別子選出部
文書ベクトル生成部
文書ベクトル記憶部
分類部
結果記憶部
205
206
207
208
(a)
209
特徴ベクトル生成用辞書
201
202
203
204
文書記憶部
文書解析部
概念ベクトル生成部
概念ベクトル記憶部
(b)
209
204
210
特徴ベクトル生成用辞書
概念ベクトル記憶部
有用概念識別子選出部
文書記憶部
文書解析部
文書ベクトル生成部
文書ベクトル記憶部
分類部
結果記憶部
201
202
205
206
207
208
(c)

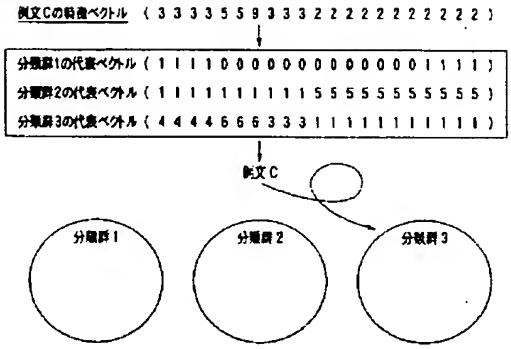
【図3】


309
特徴ベクトル生成用辞書
301
302
303
304
310
文書記憶部
文書解析部
概念ベクトル生成部
概念ベクトル記憶部
有用概念識別子選出部
文書ベクトル生成部
文書ベクトル記憶部
分類部
結果記憶部
305
306
307
308
代表ベクトル記憶部
代表ベクトル生成部
312
311
(a)
309
特徴ベクトル生成用辞書
301
302
303
304
文書記憶部
文書解析部
概念ベクトル生成部
概念ベクトル記憶部
(b)
310
309
304
有用概念識別子選出部
特徴ベクトル生成用辞書
概念ベクトル記憶部
文書記憶部
文書解析部
文書ベクトル生成部
文書ベクトル記憶部
分類部
結果記憶部
301
302
305
306
307
308
代表ベクトル記憶部
代表ベクトル生成部
312
311
(c)

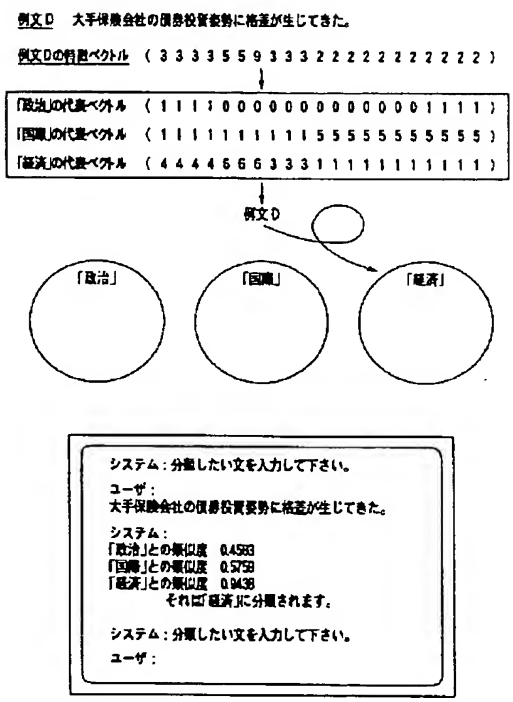
【図6】


アメリカ　（1111111111100000000000000）
政府　（1111111111100000000000000）
戦いの道具　（0000011300011111111111111）
削減　（0000011300011111111111111）
+）　提案　（1111111111100000000000000）
例文Cの特徴ベクトル　（3333559333222222222222）
例文C　「アメリカ政府は兵器の削減を提案した。」
例文Cから抽出される概念識別子
『アメリカ』『政府』『戦いの道具』『削減』『提案』

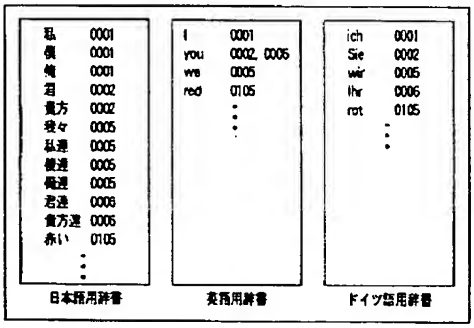
【図7】


例文Cの特徴ベクトル　（3333559333222222222222）
分類群1の代表ベクトル（1111000000000000000011111）
分類群2の代表ベクトル（1111111111155555555555555）
分類群3の代表ベクトル（4444666333111111111111111）
例文C
分類群1
分類群2
分類群3

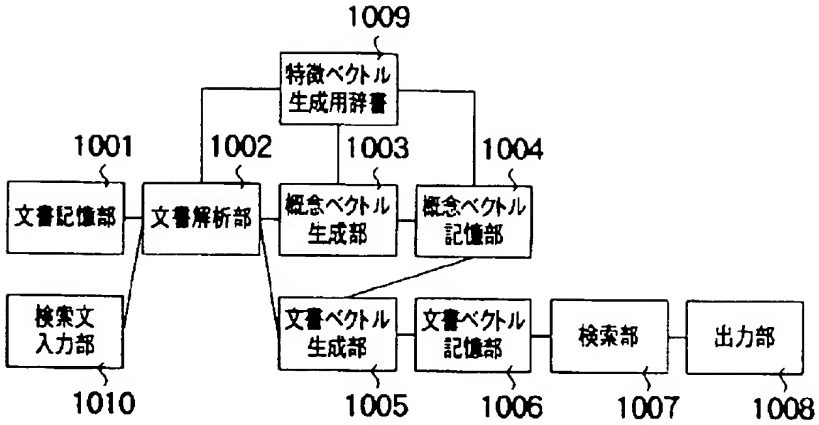
【図8】


例文D　大手保険会社の債券投資姿勢に格差が生じてきた。
例文Dの特徴ベクトル　（3333559333222222222222）
「政治」の代表ベクトル　（1111000000000000000011111）
「国際」の代表ベクトル　（1111111111155555555555555）
「経済」の代表ベクトル　（4444666333111111111111111）
例文D
「政治」
「国際」
「経済」
システム：分類したい文を入力して下さい。
ユーザ：
大手保険会社の債券投資姿勢に格差が生じてきた。
システム：
「政治」との類似度　0.4563
「国際」との類似度　0.5758
「経済」との類似度　0.9438
それは「経済」に分類されます。
システム：分類したい文を入力して下さい。
ユーザ：

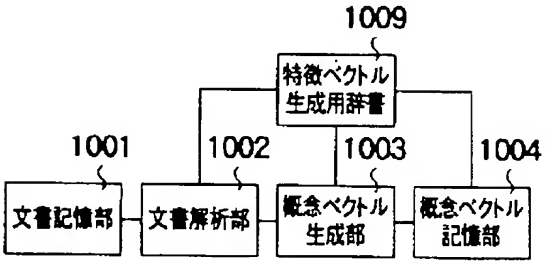
【図9】


私　0001
僕　0001
俺　0001
君　0002
貴方　0002
我々　0005
私達　0005
僕達　0005
俺達　0005
君達　0006
貴方達　0006
赤い　0105
日本語用辞書
I　0001
you　0002, 0006
we　0005
red　0105
英語用辞書
ich　0001
Sie　0002
wir　0005
Ihr　0006
rot　0105
ドイツ語用辞書

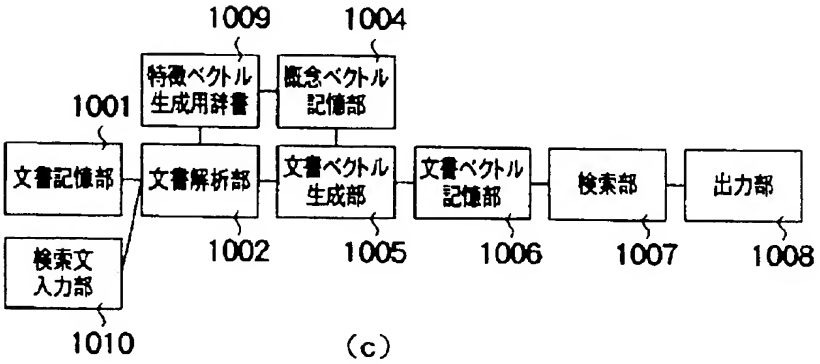
【図10】


1009
特徴ベクトル生成用辞書
1001
1002
1003
1004
文書記憶部
文書解析部
概念ベクトル生成部
概念ベクトル記憶部
検索文入力部
文書ベクトル生成部
文書ベクトル記憶部
検索部
出力部
1010
1005
1006
1007
1008

(a)

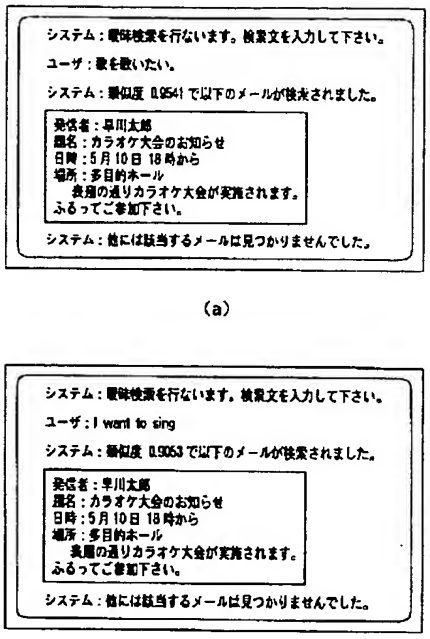

1009
特徴ベクトル生成用辞書
1001
1002
1003
1004
文書記憶部
文書解析部
概念ベクトル生成部
概念ベクトル記憶部

(b)

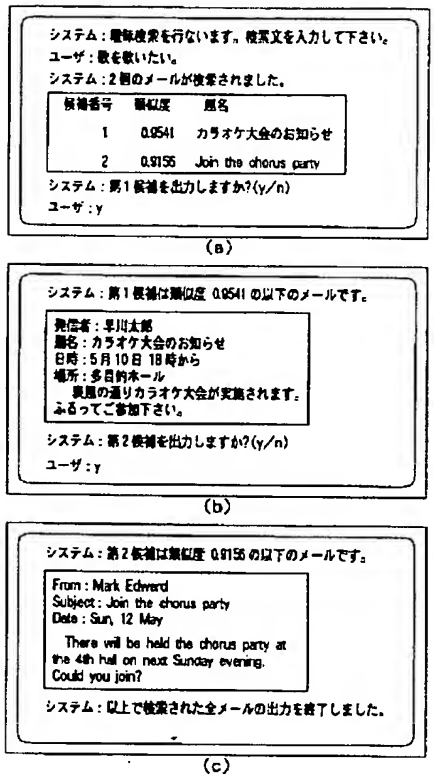

1009
1004
特徴ベクトル生成用辞書
概念ベクトル記憶部
1001
文書記憶部
文書解析部
文書ベクトル生成部
文書ベクトル記憶部
検索部
出力部
検索文入力部
1002
1005
1006
1007
1008
1010

(c)

【図１１】


システム：曖昧検索を行ないます。検索文を入力して下さい。
ユーザ：歌を歌いたい。
システム：類似度 0.9541 で以下のメールが検索されました。
発信者：早川太郎
題名：カラオケ大会のお知らせ
日時：5月10日 18時から
場所：多目的ホール
　良屋の通りカラオケ大会が実施されます。
ふるってご参加下さい。
システム：他には該当するメールは見つかりませんでした。
(a)
システム：曖昧検索を行ないます。検索文を入力して下さい。
ユーザ：I want to sing
システム：類似度 0.9053 で以下のメールが検索されました。
発信者：早川太郎
題名：カラオケ大会のお知らせ
日時：5月10日 18時から
場所：多目的ホール
　良屋の通りカラオケ大会が実施されます。
ふるってご参加下さい。
システム：他には該当するメールは見つかりませんでした。
(b)

【図１２】


システム：曖昧検索を行ないます。検索文を入力して下さい。
ユーザ：歌を歌いたい。
システム：2 個のメールが検索されました。
候補番号　類似度　題名
1　0.9541　カラオケ大会のお知らせ
2　0.9156　Join the chorus party
システム：第 1 候補を出力しますか?(y/n)
ユーザ：y
(a)
システム：第 1 候補は類似度 0.9541 の以下のメールです。
発信者：早川太郎
題名：カラオケ大会のお知らせ
日時：5月10日 18時から
場所：多目的ホール
　良屋の通りカラオケ大会が実施されます。
ふるってご参加下さい。
システム：第 2 候補を出力しますか?(y/n)
ユーザ：y
(b)
システム：第 2 候補は類似度 0.9156 の以下のメールです。
From : Mark Edward
Subject : Join the chorus party
Date : Sun, 12 May
　There will be held the chorus party at the 4th hall on next Sunday evening.
Could you join?
システム：以上で検索された全メールの出力を終了しました。
(c)